Measurement Error and Misclassification in Epidemiology Studies of Chlorpyrifos and Neurodevelopmental Outcomes

Submission to US EPA Docket # 2015-05844

April 24, 2015

GRADIENT

www.gradientcorp.com
20 University Road
Cambridge, MA 02138
617-395-5000

Introduction

In the Revised Human Health Risk Assessment for Chlorpyrifos (HHRA; US EPA, 2014), the United States Environmental Protection Agency (EPA) Office of Pesticide Programs (OPP) reviewed results from several epidemiology studies of prenatal chlorpyrifos exposure and neurodevelopmental outcomes and concluded that chlorpyrifos exposure likely played a role in observed neurodevelopmental effects. Applying principles described in its draft handbook for incorporating epidemiology data in risk assessment, OPP evaluated the potential for measurement error in available epidemiology studies and discussed its potential impacts on measured associations. OPP's evaluation was not sufficiently rigorous to provide a thorough, balanced perspective on the epidemiology literature as a whole. We have identified several specific shortcomings in OPP's assessment of measurement error in individual studies, especially pertaining to analyses of the Columbia study, the cohort to which OPP assigned the greatest weight in the overall evaluation of epidemiology evidence.

After discussing these shortcomings, we describe additional analyses that could be conducted by OPP using raw data from the Columbia study. This additional work could help resolve the remaining uncertainties regarding possible biases caused by measurement error, including residual confounding resulting from measurement error in model covariates. Finally, we present the results of two sensitivity analyses we conducted to estimate some of the potential impacts of measurement error in the Columbia study. These analyses were completed using summary data alone and are, therefore, less informative than the analyses that would be possible if we had access to raw data. Despite this, our results indicate that at least some reported associations could be biased by exposure or outcome misclassification.

Specifically, we discuss the following four points:

1. OPP has developed guidelines for evaluating potential measurement error in epidemiology studies for use in risk assessment;
2. OPP did not adhere to its own guidelines for assessing measurement error in epidemiology studies of chlorpyrifos and neurodevelopmental outcomes;
3. Unresolved uncertainties about measurement error in the Columbia study could be addressed with additional analyses of original data; and
4. Preliminary quantitative bias analyses of available summary data demonstrate that positive findings in the Columbia study could be explained by exposure or outcome misclassification.

1 OPP Has Developed Guidelines for Evaluating Potential Measurement Error in Epidemiology Studies for Use in Risk Assessment

Towards the goal of using the results of epidemiology studies in the "most scientifically robust and transparent way," OPP proposed a draft framework for weighing epidemiology results and integrating them into risk assessment (US EPA, 2010a), and the office solicited comments from the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Science Advisory Panel (SAP) as well as the general public to revise and strengthen the framework.

Several aspects of the framework are relevant to measurements of exposure, outcome, and covariates, and directly or indirectly refer to the impact of measurement error or misclassification on observed epidemiology associations.[1] Although the framework lacked specific guidance, it indicated that the most useful epidemiology studies for risk assessment employ reliable and valid exposure assessment methods. In addition, OPP stated that one should consider whether covariates relevant to confounding are properly described, measured, and analyzed. Finally, OPP acknowledged that imperfect measurements of exposure or outcome can lead to information bias, which can bias an observed association in either the positive or negative direction. The framework also stated that statistical methods should be evaluated and that epidemiology studies incorporated into risk assessment should include complete descriptions of statistical approaches. This concept applies to measurement error, because choices made in analysis can introduce or amplify biases arising from various types of measurement error, as we demonstrate with specific examples below.

The FIFRA SAP reviewed the framework and commended OPP for developing it (US EPA, 2010b). The SAP reiterated the general importance of establishing clear and robust guidelines for evaluating the quality of epidemiology data, emphasizing "the quality and reliability of the information provided by epidemiologic studies needs to be closely scrutinized" (US EPA, 2010b). To strengthen the framework, SAP recommended that OPP develop a set of specific criteria for determining the acceptability of epidemiology studies, including the use of sensitivity analyses to test the robustness of study results to measurement error in exposure as well as covariates used to adjust for confounding. In a recent publication, EPA scientists from various centers and offices, including OPP, also emphasized the importance of sensitivity analyses when applying epidemiology results to human health risk assessment (Christensen *et al.*, 2015).

In summary, even though the draft framework lacked specific guidance on how measurement error should be assessed and, importantly, how to integrate epidemiology data into risk assessment when measurement error is significant, the framework reinforced EPA's values of transparency and scientific rigor. Feedback from the SAP provided OPP with specific suggestions for incorporating quantitative methods for assessing the impacts of measurement error. Below, we review OPP's evaluation of the epidemiology data used in the chlorpyrifos risk assessment in light of OPP's draft framework and the SAP's feedback on it.

2 OPP Did Not Adhere to Its Own Guidelines for Assessing Measurement Error in Epidemiology Studies of Chlorpyrifos and Neurodevelopmental Outcomes

In the revised HHRA for chlorpyrifos, OPP reviewed 17 peer-reviewed research reports describing prenatal chlorpyrifos exposure and neurodevelopmental outcomes in three children's health cohorts. Of the three cohorts, OPP placed the greatest weight on the results of the "Columbia study," a cohort of minority women and children in New York City, because its exposure assessment was based on cord blood concentrations of chlorpyrifos. By contrast, prenatal exposure in the other two cohorts (the Mt. Sinai study, which also enrolled minority women and children in New York City, and the CHAMACOS study, which included mother-child pairs living in an agricultural community in California) was estimated based on concentrations of chlorpyrifos metabolites in maternal urine. Cord blood chlorpyrifos is considered to be a superior method for estimating prenatal chlorpyrifos exposure for a number of reasons (Prueitt *et al.*, 2011; Eaton *et al.*, 2008). Based largely on the positive associations reported in the Columbia study publications, OPP concluded that chlorpyrifos "likely" played a role in neurodevelopmental outcomes observed in the epidemiology studies.

[1] The terms "measurement error" and "misclassification" refer to errors in continuous and categorical variables, respectively, and are not completely interchangeable. For this report we will use the term "measurement error" for both types, to be concise, except in cases when it is necessary to distinguish between the two types of error.

OPP discussed the general limitations of chlorpyrifos epidemiology studies overall and further detailed the strengths and weaknesses of individual studies in Appendix 3 of the HHRA (US EPA, 2014). Limitations described by OPP included measurement error in estimation of exposures, outcome, and covariates, and, in many cases, OPP qualitatively estimated the potential impacts of error on measured associations. This type of critical evaluation is in line with the OPP framework for incorporating epidemiology results into risk assessment. However, OPP's evaluation overall lacked sufficient rigor and consistency necessary to provide an accurate, unbiased perspective on the results of the epidemiology studies reviewed in the HHRA. Below, we identify critical shortcomings in OPP's assessments of measurement error for each type of measurement (*i.e.*, exposure, outcome, and covariate), and then describe how a more rigorous approach to evaluating these errors would have increased the utility of the epidemiology results in the chlorpyrifos risk assessment.

Exposure Measurement Error

In Section 2.3 of the HHRA (US EPA, 2014), OPP described general factors leading to errors in exposure assessment in all three cohorts. OPP determined that the challenge of estimating accurate chlorpyrifos doses during the most relevant periods of development is a major limitation common to all studies. Sources of measurement error largely stem from the variability in chlorpyrifos biomarker concentrations over short time scales and the fact that exposure assessment in all studies was based on only one biomarker measurement (or, occasionally, two). For these reasons, exposure estimates used in analyses may have differed substantially from true *in utero* exposures experienced during critical developmental periods. In addition, the CHAMACOS and Mt. Sinai studies relied on concentrations of pesticide metabolites (*i.e.*, TCPy and dialkyl phosphates [DAPs]) in maternal urine for exposure assessment. Concentrations of these metabolites have relatively poor specificity for chlorpyrifos exposure, because they reflect exposure to other organophosphate pesticides as well as preformed nontoxic TCPy and DAPs in the environment (Morgan *et al.*, 2005; Lu *et al.*, 2005). Therefore, exposure estimates used in the CHAMACOS and Mt. Sinai studies were affected by additional sources of measurement error beyond those in the Columbia study.

The HHRA further expanded on aspects of measurement error in its detailed reviews of individual studies (US EPA, 2014, Appendix 3), and, in all cases, it predicted that exposure measurement error was nondifferential with respect to neurodevelopmental outcomes. OPP repeatedly stated that because the errors are nondifferential, they likely biased observed epidemiology associations towards the null, thereby masking any true relationships. While it is true that nondifferential exposure measurement error will often have this effect on measured associations, it is not always the case. Nondifferential error is guaranteed to bias associations towards the null only under specific conditions, none of which were critically assessed by OPP or individual chlorpyrifos epidemiology researchers.

Several quantitative analyses have demonstrated realistic scenarios under which approximately nondifferential exposure measurement errors can bias results away from the null (Flegal *et al.*, 1991; Jurek *et al.*, 2008; Dosemeci *et al.*, 1990). Jurek *et al.* (2008) showed that associations measured in datasets with low exposure prevalence are especially vulnerable to exposure misclassification that is nearly, but not completely, nondifferential. Overall, the possibility that exposure measurement error can bias results away from the null should not be dismissed by OPP, especially considering that it is common practice for researchers to run multiple statistical models and selectively present the results of models that yield positive findings. In fact, model selection bias is evident in the Columbia study, as discussed below.

OPP also failed to consider that the practice of categorizing continuous exposure measurements can lead to exposure misclassification. Statisticians generally discourage using categories when continuous values are available, because doing so results in lower variability and, subsequently, reduced statistical power to detect true effects (Froslie *et al.*, 2010). Another disadvantage of this approach is that categorizing continuous exposure measurements affected by nondifferential errors can lead to differential misclassification errors, potentially biasing observed associations in either direction (Flegal *et al.*, 1991). In all three children's health cohorts, continuous exposure measurements were grouped into categories for at least some analyses (Barr *et al.*, 2010; Eskenazi *et al.*, 2004, 2007; Engel *et al.*, 2007; Berkowitz *et al.*, 2004; Rauh *et al.*, 2006; Lovasi *et al.*, 2011), but OPP did not mention any potential biases associated with this approach.

In addition, there can be more serious consequences of categorizing continuous measures of exposure. A close look at the statistical methods employed in the Columbia study indicate that categorization of continuous chlorpyrifos measurements most certainly biased study findings away from the null. Rauh *et al.* (2006) estimated adjusted odds ratios (ORs) of 2.37 (95% confidence interval [CI]: 1.08-5.19) and 4.52 (95% CI: 1.61-12.70) for mental and psychomotor delay, respectively, in association with high *versus* low chlorpyrifos exposure. However, their method of defining high and low exposure groups likely contributed to false positive results. In the Methods section, the authors stated that preliminary analyses of Mental Development Index (MDI) and Psychomotor Development Index (PDI) scores indicated neither a "linear or nonlinear dose-response relationship between chlorpyrifos levels and developmental outcomes," but they provided no details about how this was determined (Rauh *et al.*, 2006).

Next, they explored associations across various categories of exposure. Continuous chlorpyrifos levels were categorized into four groups, consisting of concentrations that were less than the limit of detection (LOD) (n = 80) and tertiles of those that were detectable (*i.e.*, first tertile, n = 65; second tertile, n = 39; and third tertile, n = 44). Rauh *et al.* (2006) calculated effect estimates for each category and observed that the strongest associations resulted when exposure groups were redefined in a dichotomous manner, with low and high exposure groups defined as below and above 6.17 pg/g, respectively, *i.e.*, the concentration cut-off between the third and fourth highest categories of exposure. This description of preliminary results appeared only in the Methods section of the article, as an explanation for the choice of the 6.17 pg/g cut-point to define low *vs.* high exposure. By contrast, in the Results section of the article, Rauh *et al.* (2006) mentioned neither the null findings of their preliminary analysis nor the weaker associations observed for alternative categorization schemes.

Another source of exposure measurement error that received little attention in the HHRA was the treatment of nondetectable biomarker readings. Chlorpyrifos cord blood measurements for 43% of the Columbia cohort fell below the LOD. In Rauh *et al.* (2011), the values below the LOD were imputed so that exposure could be analyzed as a continuous variable. Based solely on an assumption of a lognormal distribution, missing values were assigned an expected value based on the distributional shape. That is, the nondetectable chlorpyrifos concentrations were assigned the most likely value expected based on the predicted shape of a log-normal distribution extending below the LOD. In statistics, this expected value is referred to as "$E(X|X < LOD)$."

The advantage of imputing nondetectable chlorpyrifos concentrations is that all subjects can be included in the analyses, and the method of imputation used by Rauh *et al.* (2001) yields unbiased regression results when the proportion of nondetectable values is small. For example, Lubin *et al.* (2004) conducted simulations to critically evaluate various methods for imputing nondetectable exposure measurements and the impacts on subsequent regression analyses. When the proportion of nondetectable values was modest (*i.e.*, 5-10%), substitution of $E(X|X < LOD)$ produced valid results. However, in datasets with larger proportions of nondetectable values, substitution of $E(X|X < LOD)$ resulted in biased coefficients and standard errors, and Lubin *et al.* (2004) concluded that more sophisticated treatments of nondetectable values, such as multiple imputation, is warranted in this scenario.

Rauh *et al.* (2011) gave little attention to the possibility that their treatment of nondetectable chlorpyrifos concentrations may have led to bias. They described a sensitivity analysis in which the regression analysis was repeated with only the detectable chlorpyrifos levels and stated that they observed "no consistent differences in estimates." No data were shown to support this statement, so it is difficult to judge the validity of this argument; also, this sensitivity analysis does not address potential biases in standard errors, which can affect inferences based on regression results. Furthermore, the investigators' conclusion about a lack of a threshold at low doses is undermined by the fact that the lower 43% of the chlorpyrifos concentrations were nondetectable and imputed using a potentially biased approach.

Outcome Measurement Error

In the HHRA, OPP discussed the challenges inherent in accurately measuring neurobehavioral outcomes in young children and acknowledged potential biases that could have affected the results of the studies they reviewed. A single clinical neurobehavioral test result in an individual is not sufficient to consider an outcome as a functional impairment or illness, even if the result is "abnormal." Also, studies assessing single measurements of neurobehavioral outcome at each time point may be subject to additional measurement error as a result of within-subject variability in results (Eaton *et al.*, 2008). There are many extraneous factors that affect the results of clinical tests, such as test administrator training and blinding to exposure status, and the child's physical activity level, diet, medication use, co-exposures, and whether or not they are obese (Eaton *et al.*, 2008). In particular, many of these tests require an advanced level of training and expertise in test administration (Leonard *et al.*, 2001). Taken together, these factors indicate that outcome measurements may have been highly influenced by errors and misclassification.

Despite several limitations in assessment methods, both OPP and the SAP concluded that the chlorpyrifos epidemiology studies they reviewed utilized the "best available" measurement tools and conducted testing in consistent and standardized ways. OPP predicted that most measurement errors in outcome assessment were nondifferential and that, as a result, any bias in the measured epidemiology results likely would have been towards the null. Our previous discussion regarding nondifferential measurement error applies by analogy to outcome measurement as well. Specifically, nondifferential outcome error does not guarantee that bias is towards the null, except under very specific conditions. If outcome error is nearly, but not perfectly, nondifferential in nature, the error or misclassification can bias associations in either direction, and studies with rare outcomes are especially susceptible to this (Jurek *et al.*, 2008). In the case of the three binary outcomes related to behavioral disorders at 36 months analyzed in Rauh *et al.* (2006), a very small number of children were diagnosed as having a behavioral problem, and any outcome misclassification could have biased associations substantially. We explore this possibility in a quantitative bias analysis in Section 4, below.

In addition, several continuous measures of neurodevelopmental outcomes were dichotomized for use in logistic regression, and the choice of cut-points for diagnosing delayed *versus* non-delayed children may have strongly influenced results. In the Columbia study, scores of 85 on the PDI and the MDI were used to distinguish between children who were "normal" *versus* "delayed," but no rationale or citation for this specific cut-point was provided. In contrast, other sources indicate that the typical cut-offs for moderate and severe development delay using the BSID-II are 70 and 55, respectively (Bos, 2013). In the absence of justification for a cut-off of 85, it is plausible that the categories have been defined to maximize positive findings, as was the case for exposure categories, described above.

Finally, OPP briefly noted that differential errors are possible in one outcome assessment tool, the Child Behavior Checklist (CBCL). Aside from a brief mention of this possibility, however, OPP did not discuss the potential impact that utilizing this tool may have had on study findings (Eskenazi *et al.*, 2006, 2007;

Rauh *et al.*, 2006). The CBCL is a survey completed by mothers and is based on subjective judgment of child behavior. It is feasible that mothers would be more or less likely to report behavioral issues based on study results at earlier time points in the follow-up period. For example, mothers who tested high for chlorpyrifos or other chemicals at the initiation of the study may have been more likely to suspect that this exposure could be the cause of behavior disorders and, therefore, may have been more likely to over-report problematic symptoms their child developed at later time points.

If mothers of children in higher exposure categories differentially over-reported child symptoms at age 36 months on the CBCL, effect estimates would have been biased high. Given the very small number of children identified as having behavioral problems based on the CBCL in Rauh *et al.* (2006), the impact of only one or two misclassified children could have had a profound impact on the measured associations. Rauh *et al.* (2006) did not present counts of children with and without behavioral problems in each exposure group, but Table 3 in their paper shows that only 3.4% of the cohort was diagnosed with attention problems at 36 months. This indicates that there were seven children diagnosed with attention problems, out of the cohort of 228 children. This small number of children in the clinical range for attention problems is reflected in the very wide CIs calculated in the multivariate logistic regression (OR = 11.26, 95% CI: 1.79-70.99 for attention problems) and indicates the risk estimate is not stable.

Covariate Measurement Error

The HHRA noted that several important confounding factors may have biased the results of the chlorpyrifos epidemiology studies OPP reviewed in either positive or negative directions. The potential for confounding in these studies is especially high because several maternal characteristics are strongly associated with both exposure and outcome. For example, as acknowledged by OPP in the HHRA, ample research has established that early life neurodevelopment is positively associated with indicators of increased socioeconomic status (SES). Quantitative analyses have demonstrated that epidemiology studies of low dose environmental exposures and neurodevelopmental outcomes can be confounded by maternal intelligence, home environment, and SES, even if differences in these factors between exposure groups are small (Mink *et al.*, 2004). Mink *et al.* (2004) found that substantial confounding can occur even when these variables are measured and included as adjustment variables, because measures of SES are often inaccurate and residual confounding may persist in multivariate regression.

OPP claims that the issue of confounding was addressed, in part, by the restriction of study cohorts to relatively homogeneous populations. However, women enrolled in each of the three cohort studies displayed a substantial amount of heterogeneity in several ways. For example, significant associations between outcomes and multiple maternal characteristics, including environmental tobacco smoke (ETS) exposure, material hardship, and maternal IQ were observed within the Columbia study cohort, demonstrating that confounding was likely (Rauh *et al.*, 2004, 2011). Likewise, in the CHAMACOS cohort, higher DAPs were measured in mothers with lower intelligence and lower HOME scores (*i.e.*, lower measures of the quality of care-taking environment) (Bouchard *et al.*, 2011).

Even though all the chlorpyrifos epidemiology studies employed multivariate analyses to mitigate confounding, measurement error in covariates limits the effectiveness of the statistical adjustments. Maternal smoking, drug use, and drinking during pregnancy were ascertained by self-report, and these stigmatized behaviors are likely under-reported by many women. It is striking that, despite this, the prevalence of drinking during pregnancy in the Columbia cohort was estimated to be 25% (Whyatt *et al.*, 2004), but none of the Columbia study analyses considered confounding by alcohol use. In fact, this maternal behavior was not mentioned in any reports that followed Whyatt *et al.* (2004). OPP indirectly addressed this topic by noting that a small number of women used alcohol in the Columbia cohort and citing the low percentage of women who reported engaging in "heavy drinking." Because a substantial proportion of the Columbia cohort reported drinking, and prenatal alcohol exposure may confound the relationship between chlorpyrifos and outcomes, reported associations may have been biased.

In addition, data on some covariates were collected as continuous measures but then dichotomized for use in multivariate analysis, thereby artificially reducing the variability and effectiveness of statistical adjustment (Rauh *et al.*, 2006; 2011; Eskenazi *et al.*, 2007). This practice is commonly discouraged by statisticians (Altman, 2006), and this is an important limitation of the Columbia study in particular.

Covariate measurement error is also likely to have limited researchers' ability to determine what factors confound epidemiology associations. Columbia study researchers asserted that certain factors could not confound relationships on the basis of statistical significance testing. For example, they concluded that because correlations between blood lead levels and both exposure and outcome were not significant, lead was not a confounder; OPP agreed with this assessment. However, correlation was assessed in a subsample of only 89 mother-child pairs, and the test was likely underpowered to detect true associations. Neither OPP nor the researchers considered that small sample sizes and measurement error in covariates limited the statistical power to detect true associations. OPP discussed limited samples sizes and measurement error elsewhere in the HHRA, but only as factors that may have masked true associations. In contrast, the HHRA generally did not give attention to the methodological limitations that may have had the opposite effect.

This use of statistical significance testing for assessing whether a certain factor acts as a confounder and, likewise, for selecting covariates to include in adjusted models is generally discouraged by epidemiologists (Rothman *et al.*, 2008). Interestingly, OPP noted this several times in the HHRA in reviews of individual studies, but only applied the criticism to the CHAMACOS and Mt. Sinai studies. Columbia study analyses should be held to the same standard; OPP should more closely scrutinize the decisions researchers made regarding model covariates and the resulting potential for residual confounding.

Shortcomings in OPP's Overall Summary of Measurement Errors in Chlorpyrifos Epidemiology Studies

In the HHRA's overall integration of epidemiology evidence, the relative impacts of exposure measurement error and residual confounding were directly compared. OPP concluded that, even though it is possible that residual confounding biased observed epidemiology associations away from the null, this bias was likely weaker than the effects of nondifferential exposure measurement error, which OPP maintained biased results towards the null. OPP indicated that this argument was supported by a review of exposure measurement error and confounding in occupational epidemiology studies (Blair *et al.*, 2007). However, Blair *et al.* (2007) is not directly applicable to chlorpyrifos epidemiology research, because exposure assessment in occupational settings typically involves record reviews or retrospective self-reports; these methods are much more susceptible to measurement errors, including differential errors such as recall bias. In contrast, the chlorpyrifos epidemiology studies that OPP reviewed in the risk assessment depended on objective biomarker measurements of exposure, which are far less susceptible to differential measurement errors.

More careful consideration of the magnitude and implications of exposure, outcome, and covariate measurement errors in these studies is needed. As we discussed in depth above, there are many ways by which measurement error, even when nondifferential, may have contributed to false positive findings between chlorpyrifos and neurodevelopmental outcomes. Without additional evaluation, OPP's conclusion that chlorpyrifos exposure likely played a role in neurodevelopmental effects is not well supported.

3 Unresolved Uncertainties About Measurement Error in the Columbia Study Could Be Addressed with Additional Analyses of Original Data

Following two SAP reviews of the draft chlorpyrifos HHRA, OPP determined that certain areas of uncertainty limited the incorporation of Columbia study results into the risk assessment (US EPA, 2008, 2012). OPP therefore requested that Columbia study researchers provide the original analytic file used to conduct the analyses reported in Rauh *et al.* (2006; 2011) and Whyatt *et al.* (2004), so that uncertainties could be carefully evaluated. Columbia researchers refused the request, but agreed to meet with OPP researchers to address OPP's concerns. Based on the discussion at this meeting and additional information subsequently provided to OPP on request, OPP dropped its previous request for original data (US EPA, 2014, Appendix 6, p. 384).

However, several key uncertainties were inadequately addressed by the additional information and analyses, and OPP should renew the request to access the studies' original raw data. In this section, we describe several analyses that should be conducted using the raw data to address the remaining uncertainties about measurement error in the Columbia study. Doing so would help OPP achieve its goal of transparency while also critically evaluating epidemiology data utilized in the risk assessment. This is especially important for the Columbia study, given the greater weight it received in the HHRA evaluation.

None of the analyses we suggest below require additional data collection and most are simple to perform, with results that are easy to interpret. Our final suggestion is a more in-depth analysis aimed at accounting for multiple errors and uncertainties simultaneously. This type of quantitative bias assessment would be more difficult to conduct and interpret, but the results would provide critical insights into the potential impact of errors on the chlorpyrifos epidemiology studies that OPP evaluated. Researchers in academia, government, and industry have called for an increased use of such methods to improve the utility of epidemiology data in human health risk assessment (Burns *et al.*, 2014).

Methods of Adjusting for Potential Confounders

OPP maintained that Columbia study researchers addressed the potential for confounding "to the extent possible," but we believe that additional analyses could provide meaningful insight into the magnitude of residual confounding in reported associations. Simple analyses should be conducted to test the assertion of both OPP and Columbia investigators that lead, polycyclic aromatic hydrocarbon (PAH), alcohol, and ETS exposure did not confound the positive associations between chlorpyrifos and neurodevelopment. Rather than rely on correlational analyses and the results of statistical significance testing, the main findings of the Columbia study should be re-analyzed to determine whether results are sensitive to the inclusion of lead levels, PAH exposures, and reported maternal drinking. Lead exposure data were available for only a subset of children, but missing values could be imputed fairly easily using other available characteristics of mothers and children.

Similarly, additional analyses should explore whether factors that have been dichotomized for multivariate analysis (*e.g.*, years of maternal education and household income) have a stronger impact on measured associations when included in the model as continuous variables.

Treatment of Missing Data

OPP should evaluate whether the method used to impute missing chlorpyrifos measurements in Rauh *et al.* (2011) could have biased regression coefficients or affected the size of standard errors. With the original dataset, OPP could assess whether the results of regression are sensitive to variations on the imputation method employed. Simulations conducted by Lubin *et al.* (2004) showed that the regression results were highly sensitive to the methods used to impute exposures below the LOD, especially when the proportion of missing data is relatively large, as is the case in the Columbia study.

Ad hoc Cut-points Chosen for Categorization of Exposure, Outcome, and Covariate Variables

As described above, the Columbia researchers used inappropriate methods to define high- and low-exposure groups for analyses of several neurobehavioral outcomes (Rauh *et al.*, 2006). Similarly, cut-points for the dichotomization of outcomes from continuous measurements appeared to be arbitrary. In fact, it is possible that these two potential sources of bias were compounded in the analyses of dichotomized exposures and outcomes. The small numbers of cases in high-exposure categories (see Table 1) increases the likelihood that minor variations in the cut-points could have substantial impacts on the counts of exposed cases and noncases and, subsequently, on measured associations (Jurek *et al.*, 2008).

To rigorously assess whether key epidemiology findings from the Columbia study are sensitive to cut-points, further analyses should be conducted in which main effects are recalculated for a variety of exposure and outcome category cut-points. An analogous set of sensitivity analyses focused on the cut-points applied to dichotomized confounders, such as years of maternal education, should be conducted as well.

Variations in Subject Characteristics Before and After the Chlorpyrifos Ban

In the Columbia study, exposures to chlorpyrifos dramatically decreased across the 6-year period during which enrolled mothers gave birth. As a result, chlorpyrifos exposure is strongly correlated with calendar time in this cohort. If characteristics of the enrolled subjects varied over time, a false association between chlorpyrifos and neurobehavioral outcomes could have occurred. If recruiting strategies or locations changed across the 6 years of subject enrollment, for example, it is plausible that women recruited later in the study period were consistently higher or lower in SES or some other factor strongly associated with child neurodevelopment. Even though researchers attempted to control for SES-related factors in their analyses, residual confounding was likely to have occurred, for the reasons discussed above. Close inspection of maternal characteristics and patterns over time may indicate that the characteristics of enrolled women shifted over the recruitment period, and this should prompt increased scrutiny on the methods used to control for confounding.

Probabilistic Bias Assessment of Multiple Measurement Errors and Biases

Finally, a detailed quantitative assessment of potential biases should be conducted with the original data from the Columbia study. As an alternative to the deterministic approaches to assessing potential biases one-by-one, described above, probabilistic methods could be employed to explore the impact of exposure and outcome measurement error, selection bias, and unmeasured confounding simultaneously. Use of the original datasets would ensure that the correlation structure is preserved. For this analysis, estimates of the magnitude of differential and nondifferential errors in exposure and outcome measures are assumed, and uncertainty in these parameters is modeled using distributions of plausible values. Then, Monte Carlo sampling of parameters from probability distributions and reanalysis of the dataset over thousands of iterations produces distributions of effect estimates that reflect uncertainty, bias, and variability in measured associations. Several examples of this approach can be found in the literature, and researchers have called for increased utilization of these methods in epidemiology studies (Maldonado, 2008; Lash and Fink, 2003; Meliker *et al*., 2010; Lash, 2009).

4 Preliminary Quantitative Bias Analyses of Available Summary Data Demonstrate That False Positive Findings in the Columbia Study Could Be Explained by Exposure or Outcome Misclassification

In the absence of original individual-level data, quantitative bias analyses can be conducted using summary data of the type generally presented in research publications (Lash, 2009). Even though this type of sensitivity analysis is less informative than that which is possible with individual-level data, these analyses can provide important insights into the potential ramifications of measurement error, selection bias, and/or unmeasured confounding on reported associations. We provide two examples of deterministic sensitivity analyses below, both of which were conducted using the summary results provided in Rauh *et al.* (2006). The first is a deterministic analysis illustrating how the choice of exposure cut-points is highly influential on estimated ORs for psychomotor and mental delay associated with various categories of chlorpyrifos exposure. As we discussed above, this variability in ORs may have led to false positive findings if researchers chose cut-points to maximize associations. The second analysis is a deterministic analysis of the potential impact of differential outcome misclassification on reported relationships with CBCL-derived outcomes in Rauh *et al.* (2006).

Sensitivity of Epidemiology Results to Variations in Exposure Categorization

Rauh *et al.* (2006) reported an adjusted OR of 2.37 (95% CI: 1.08-5.19) for mental delay and an adjusted OR of 4.52 (95% CI: 1.61-12.70) for psychomotor delay in association with high *versus* low chlorpyrifos exposure. In response to a request from the FIFRA SAP, the authors provided a more detailed breakdown of cases and noncases across four exposure categories (Table 1).

Using these data, we have calculated crude ORs to explore the variation in ORs observed using four separate schemes of exposure categorization.[2] Calculation of adjusted OR is not possible without raw data, so we have conducted this analysis with unadjusted risk estimates; we expect that similar patterns would result in multivariate analyses as well. As shown in Table 2 and 3, the magnitude and precision of results is highly sensitive to methods used to model exposure. The interpretation of the main findings of Rauh *et al.* (2006) is impacted by this inconsistency in risk estimates across categorization schemes as well as the investigators' selective presentation of the strongest and most precise risk estimates in their publication.

Analysis of Bias from Outcome Misclassification

We conducted a quantitative bias assessment to determine whether risk estimates for the three CBCL-derived outcomes (attention problems,[3] attention deficit hyperactivity disorder [ADHD] problems, and

[2] ORs were calculated in the traditional method using a 2 x 2 table: OR= (# diseased in exposed group)*(# nondiseased in unexposed group)/[(# diseased in unexposed group)*(# diseased in unexposed group)].

[3] Rauh *et al.* (2006) described outcomes as attention, ADHD and PDD "problems" based on the 98th percentile of CBCL scores in each domain, and we use the same wording here.

pervasive developmental disorder [PDD]) reported by Rauh *et al.* (2006) could be affected by outcome misclassification. As we discussed above, a small number of children were classified as having each problem at 36 months of age, ranging from 7 children identified as having attention problems to 11 with PDD problems. Because CBCL results are based on subjective reports of mothers and because parents of children found to be highly exposed to chlorpyrifos during pregnancy may be more likely to report health problems, it is feasible that some exposed children were misclassified as having a behavioral problem.

To explore the impact of misclassification of one, two, or three exposed children on risk estimates, we reconstructed 2 x 2 tables for CBCL outcomes based on percentages presented in Table 3 of Rauh *et al.* (2006). Only adjusted ORs were presented in the paper, but we did not have access to individual-level data and, so, conducted our analysis instead using crude ORs in the same method described above. We expect that impacts of misclassification would be similar in an adjusted analysis.

As shown in Table 4, misclassification of a small number of children in the exposed category would have a dramatic impact on the magnitude of risk estimates for these three outcomes. This indicates that the risk estimates reported by Rauh *et al.* (2006) may be vulnerable to outcome misclassification, even if only 1 or 2 children out of the 228 in the cohort were falsely identified as having a behavioral problem based on subjective parental reports. It is important to note that misclassification of outcome may have been differential, or it may have been nondifferential, but by random chance affected the high exposure category specifically. Regardless of the mechanism, the impact of a very small amount of misclassification would be substantial.

5 Conclusion

OPP established a framework for incorporating epidemiology research into risk assessment, which includes an evaluation of the accuracy of exposure, outcome, and covariate measurements. In the chlorpyrifos HHRA, OPP critically assessed the nature and magnitude of errors in these measurements for individual epidemiology studies, as well as for the body of research as a whole. However, the HHRA's evaluation of the impact of these errors was not sufficiently rigorous or consistent. OPP could greatly improve the utility of the Columbia study epidemiology results for risk assessment by conducting quantitative bias analyses such as those we described in Section 4 or by renewing the request for original data from investigators to pursue more substantial bias analyses, as we described in Section 3. As other researchers have noted, it is crucial to conduct quantitative sensitivity analyses when important policy decisions are to be based on the results of epidemiology research (Jurek *et al.*, 2008; Christensen *et al.*, 2015; Burns *et al.*, 2014).

Acknowledgment

This work was funded by Dow AgroSciences, LLC. This work represents the individual professional views of the authors and not necessarily the views of Dow AgroSciences, LLC.

References

Altman, DG; Royston, P. 2006. "The cost of dichotomising continuous variables." *BMJ* 332(7549):1080. doi: 10.1136/bmj.332.7549.1080.

Barr, DB; Ananth, CV; Yan, X; Lashley, S; Smulian, JC; Ledoux, TA; Hore, P; Robson, MG. 2010. "Pesticide concentrations in maternal and umbilical cord sera and their relation to birth outcomes in a population of pregnant women and newborns in New Jersey." *Sci. Total Environ.* 408(4):790-795.

Berkowitz, GS; Wetmur, JG; Birman-Deych, E; Obel, J; Lapinski, RH; Godbold, JH; Holzman, IR; Wolff, MS. 2004. "*In utero* pesticide exposure, maternal paraoxonase activity, and head circumference." *Environ. Health Perspect.* 112(3):388-391.

Blair, A; Stewart, P; Lubin, JH; Forastiere, F. 2007. "Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures."*Am. J. Ind. Med.* 50(3):199-207.

Bos, AF. 2013. "Bayley-II or Bayley-III: What do the scores tell us (Commentary)?" *Dev. Med. Child Neurol.* 55(11):978-979. doi: 10.1111/dmcn.12234.

Bouchard, MF; Chevrier, J; Harley, KG; Kogut, K; Vedar, M; Calderon, N; Trujillo, C; Johnson, C; Bradman, A; Barr, DB; Eskenazi, B. 2011. "Prenatal exposure to organophosphate pesticides and IQ in 7-year old children." *Environ. Health Perspect.* 119(8):1189-1195.

Burns, CJ; Wright, JM; Pierson, JB; Bateson, TF; Burstyn, I; Goldstein, DA; Klaunig, JE; Luben, TJ; Mihlan, G; Ritter, L; Schnatter, AR; Symons, JM; Yi, KD. 2014. "Evaluating uncertainty to strengthen epidemiologic data for use in human health risk assessments." *Environ. Health Perspect.* 122(11):1160-1165. doi: 10.1289/ehp.1308062.

Christensen, K; Christensen, CH; Wright, JM; Galizia, A; Glenn, BS; Scott, CS; Mall, JK; Bateson, TF; Murphy, PA; Cooper, GS. 2014. "The use of epidemiology in risk assessment: Challenges and opportunities." *Hum. Ecol. Risk Assess.* doi: 10.1080/10807039.2014.967039.

Dosemeci, M; Wacholder, S; Lubin, JH. 1990. "Does nondifferential misclassification of exposure always bias a true effect toward the null value?" *Am. J. Epidemiol.* 132(4):746-748.

Eaton DL; Daroff RB; Autrup H; Bridges J; Buffler P; Costa LG; Coyle J; McKhann G; Mobley WC; Nadel L; Neubert D; Schulte-Hermann R; Spencer PS. 2008. "Review of the toxicology of chlorpyrifos with an emphasis on human exposure and neurodevelopment." *Crit. Rev. Toxicol.* 38(Suppl. 2):1-125.

Engel, SM; Berkowitz, GS; Barr, DB; Teitelbaum, SL; Siskind, J; Meisel, SJ; Wetmur, JG; Wolff, MS. 2007. "Prenatal organophosphate metabolite and organochlorine levels and performance on the Brazelton Neonatal Behavioral Assessment Scale in a multiethnic pregnancy cohort." *Am. J. Epidemiol.* 165(12):1397-404.

Eskenazi, B; Marks, AR; Bradman, A; Harley, K; Barr, DB; Johnson, C; Morga, N; Jewell, NP. 2007. "Organophosphate pesticide exposure and neurodevelopment in young Mexican-American children." *Environ. Health Perspect.* 115(5):792-798.

Eskenazi, B; Harley, K; Bradman, A; Weltzien, E; Jewell, NP; Barr, DB; Furlong, CE; Holland, NT. 2004. "Association of in utero organophosphate pesticide exposure and fetal growth and length of gestation in an agricultural population." *Environ. Health Perspect.* 112(10):1116-1124.

Eskenazi, B; Marks, AR; Bradman, A; Fenster, L; Johnson, C; Barr, DB; Jewell, NP. 2006. "*In utero* exposure to dichlorodiphenyltrichloroethane (DDT) and dichlorodiphenyldichloroethylene (DDE) and neurodevelopment among young Mexican American children." *Pediatrics* 118(1):233-241.

Flegal, KM; Keyl, PM; Nieto, FJ. 1991. "Differential misclassification arising from nondifferential errors in exposure measurement." *Am. J. Epidemiol.* 134(10):1233-1244.

Froslie, KF; Roislien, J; Laake, P; Henriksen, T; Qvigstad, E; Veierod, MB. 2010. "Categorisation of continuous exposure variables revisited. A response to the Hyperglycaemia and Adverse Pregnancy Outcome (HAPO) Study." *BMC Med. Res. Methodol.* 10:103. doi: 10.1186/1471-2288-10-103.

Jurek, AM; Greenland, S; Maldonado, G. 2008.\ "How far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null?" *Int. J. Epidemiol.* 37(2):382-385.

Lash, TL; Fink, AK. 2003. "Semi-automated sensitivity analysis to assess systematic errors in observational data." *Epidemiology* 14(4):451-458. doi: 10.1097/01.EDE.0000071419.41011.cf.

Lash, TL; Fox, MP; Fink, AK. 2009. *Applying Quantitative Bias Analysis to Epidemiologic Data.* Springer, New York, NY. 192p.

Leonard, CH; Piecuch, RE; Cooper, BA. 2001. "Use of the Bayley Infant Neurodevelopmental Screener with low birth weight infants." *J. Pediatr. Psychol.* 26(1):33-40. doi: 10.1093/jpepsy/26.1.33.

Lovasi, GS; Quinn, JW; Rauh, VA; Perera, FP; Andrews, HF; Garfinkel, R; Hoepner, L; Whyatt, R; Rundle, A. 2011. "Chlorpyrifos exposure and urban residential environment characteristics as determinants of early childhood neurodevelopment." *Am. J. Public Health* 101(1):63-70. doi: 10.2105/AJPH.2009.168419.

Lu, C; Bravo, R; Caltabiano, LM; Irish, RM; Weerasekera, G; Barr, DB. 2005. "The presence of dialkylphosphates in fresh fruit juices: Implication for organophosphorus pesticide exposure and risk assessments." *J. Toxicol. Environ. Health A* 68(3):209-227. doi: 10.1080/15287390590890554.

Lubin, JH; Colt, JS; Camann, D; Davis, S; Cerhan, JR; Severson, RK; Bernstein, L; Hartge, P. 2004 "Epidemiologic evaluation of measurement data in the presence of detection limits." *Environ. Health Perspect.* 112(17):1691-1696.

Maldonado, G. 2008. "Adjusting a relative-risk estimate for study imperfections." *J. Epidemiol. Community Health* 62(7):655-663. doi: 10.1136/jech.2007.063909.

Meliker, JR; Goovaerts, P; Jacquez, GM; Nriagu, JO. 2010. "Incorporating individual-level distributions of exposure error in epidemiologic analyses: An example using arsenic in drinking water and bladder cancer." *Ann. Epidemiol.* 20(10):750-758. doi: 10.1016/j.annepidem.2010.06.012.

Mink, PJ; Goodman, M; Barraj, LM; Imrey, H; Kelsh, MA; Yager, J. 2004. "Evaluation of uncontrolled confounding in studies of environmental exposures and neurobehavioral testing in children." *Epidemiology* 15(4):385-395.

Morgan, MK; Sheldon, LS; Croghan, CW; Jones, PA; Robertson, GL; Chuang, JC; Wilson, NK; Lyu, CW. 2005. "Exposures of preschool children to chlorpyrifos and its degradation product 3,5,6-trichloro-2-pyridinol in their everyday environments." *J. Expo. Anal. Environ. Epidemiol.* 15(4):297-309.

Prueitt, RL; Goodman, JE; Bailey, LA; Rhomberg, LR. 2011. "Hypothesis-based weight-of-evidence evaluation of the neurodevelopmental effects of chlorpyrifos." *Crit. Rev. Toxicol.* 41(10):822-903.

Rauh, V; Arunajadai, S; Horton, M; Perera, F; Hoepner, L; Barr, DB; Whyatt, R. 2011. "7-Year neurodevelopmental scores and prenatal exposure to chlorpyrifos, a common agricultural pesticide. *Environ. Health Perspect.* 119(8):1196-1201.

Rauh, VA; Garfinkel, R; Perera, FP; Andrews, HF; Hoepner, L; Barr, DB; Whitehead, R; Tang, D; Whyatt, RW. 2006. "Impact of prenatal chlorpyrifos exposure on neurodevelopment in the first 3 years of life among inner-city children." *Pediatrics* 118(6):e1845-e1859.

Rauh, VA; Whyatt, RM; Garfinkel, R; Andrews, H; Hoepner, L; Reyes, A; Diaz, D; Camann, D; Perera, FP. 2004. "Developmental effects of exposure to environmental tobacco smoke and material hardship among inner-city children." *Neurotoxicol. Teratol.* 26(3):373-385.

Rothman, KJ; Greenland, S; Lash, TL. 2008. *Modern Epidemiology (Third Edition).* Lippincott Williams & Wilkins, Philadelphia, PA. 758p.

US EPA. 2012. "A Set of Scientific Issues Being Considered by the Environmental Protection Agency Regarding Chlorpyrifos Health Effects: Minutes of the FIFRA Science Advisory Panel Meeting held on April 10-12, 2012." FIFRA Scientific Advisory Panel. SAP Minutes No. 2012-04. 108p.

US EPA. 2008. "A Set of Scientific Issues Being Considered by the Environmental Protection Agency Regarding: The Agency's Evaluation of the Toxicity Profile of Chlorpyrifos: Minutes of the FIFRA Science Advisory Panel Meeting held on September 16-18, 2008." FIFRA Scientific Advisory Panel. SAP Minutes No. 2008-04. 80p., December 17. Accessed at http://www.epa.gov/scipoly/sap/meetings/2008/091608_mtg.htm.

US EPA. 2010b. "A Set of Scientific Issues Being Considered by the Environmental Protection Agency Regarding: Draft Framework and Case Studies on Atrazine, Human Incidents, and the Agricultural Health Study: Incorporation of Epidemiology and Human Incident Data into Human Health Risk Assessment." FIFRA Scientific Advisory Panel (SAP). SAP Minutes No. 2010-03. FIFRA Scientific Advisory Panel Meeting, held February 2-4 at the Environmental Protection Agency Conference Center, Arlington, Virginia. 87p., February.

US EPA. 2010a. "Framework for Incorporating Human Epidemiologic & Incident Data in Health Risk Assessment (Draft)." Office of Pesticide Programs. 68p., January 7.

US EPA. 2014. "Chlorpyrifos: Revised Human Health Risk Assessment for Registration Review." Office of Chemical Safety and Pollution Prevention. EPA-HQ-OPP-2008-0850-0195. 531p., December 29.

Whyatt, RM; Camann, D; Perera, FP; Rauh, VA; Tang, D; Kinney, PL; Garfinkel, R; Andrews, H; Hoepner, L; Barr, DB. 2005. "Biomarkers in assessing residential insecticide exposures during pregnancy and effects on fetal growth." *Toxicol. Appl. Pharmacol.* 206(2):246-254.

Whyatt, RM; Rauh, V; Barr, DB; Camann, DE; Andrews, HF; Garfinkel, R; Hoepner, LA; Diaz, D; Dietrich, J; Reyes, A; Tang, D; Kinney, PL; Perera, FP. 2004. "Prenatal insecticide exposures and birth weight and length among an urban minority cohort." *Environ. Health Perspect.* 112(10):1125-1132.

Table 1 Columbia Study Subjects Defined as Developmentally Delayed Based on MDI and PDI Scores Measured at 36 Months of Age in Each of Four Chlorpyrifos Exposure Groups Designated by Rauh *et al.* (2006)[a] These data were used in the sensitivity analyses summarized in Tables 2 and 3.

Outcome		**Group 1**	**Group 2**	**Group 3**	**Group 4**
		< LOD (n = 80)	**1st Tertile > LOD (n = 65)**	**2nd Tertile > LOD (n = 38 or 39)[b]**	**3rd Tertile > LOD (n = 45 or 44)[b]**
PDI	Psychomotor delay	7	3	3	11
	No delay	73	62	35	34
MDI	Mental delay	30	14	11	20
	No delay	50	51	28	24

Notes:

LOD = Limit of Detection; MDI = Mental Development Index; PDI = Psychomotor Development Index.

(a) Table adapted from US EPA (2014, Appendix 2, Attachment 1). Psychomotor and mental delay defined as PDI and MDI scores ≤ 85 by Rauh *et al*. (2006).

(b) The number of subjects in the two highest tertiles of exposure was inconsistent between the two outcomes.

Table 2 Odds Ratio for "Mental Delay" Calculated for Various Chlorpyrifos Exposure Categorization Schemes (Rauh *et al.*, 2006)[a] Rauh *et al.* (2006) presented results only for the exposure categorization scheme that maximized associations with mental delay. As shown here, associations calculated using a variety of other categorization schemes demonstrate that their choice of defining dichotomous exposure groups likely biased results away from the null.

Exposure Categorization	Group 1	Group 2	Group 3	Group 4
	< LOD	1st Tertile > LOD	2nd Tertile > LOD	3rd Tertile > LOD
Groups 2, 3, 4 (high) *vs.* Group 1 (low)	Reference	0.73 (0.41, 1.29)		
Groups 3, 4 (high) *vs.* Groups 1, 2 (low)	Reference		1.37 (0.78, 2.42)	
Group 4 (high) *vs.* Groups 1, 2, 3 (low)[b]	Reference			**1.95 (1.00, 3.83)**
Groups 2, 3, 4 Individually (high) *vs.* Group 1 (low)	Reference	**0.46 (0.22, 0.96)**	0.65 (0.29, 1.50)	1.39 (0.66, 2.93)
Trend Across Four Dose Groups	Linear trend p = 0.49 OR = 1.09 (0.85, 1.40) for each increase in category			

Notes:

LOD = Limit of Detection; OR = Odds Ratio.

(a) Crude ORs for "Mental Delay" (*i.e.*, MDI score ≤ 85) were calculated using counts of subjects in Table 1. Statistically signification associations at a 95% confidence level are highlighted in bold.

(b) The categorization scheme used by Rauh *et al.* (2006).

Table 3 Odds Ratio for "Psychomotor Delay" Calculated for Various Chlorpyrifos Exposure Categorization Schemes (Rauh *et al.*, 2006)[a] Rauh *et al*. (2006) presented results only for the exposure categorization scheme that maximized associations with psychomotor delay. As shown here, associations calculated using a variety of other categorization schemes demonstrate that the choice of defining dichotomous exposure groups likely biased results away from the null.

Exposure Categorization	**Group 1**	**Group 2**	**Group 3**	**Group 4**
	< LOD	**1st Tertile > LOD**	**2nd Tertile > LOD**	**3rd Tertile > LOD**
Groups 2, 3, 4 (high) *vs*. Group 1 (low)	Reference	1.35 (0.54, 3.41)		
Groups 3, 4 (high) *vs*. Groups 1, 2 (low)	Reference		**2.73 (1.16, 6.48)**	
Group 4 (high) *vs*. Groups 1, 2, 3 (low)[b]	Reference			**4.23 (1.75, 10.2)**
Groups 2, 3, 4 Individually (high) *vs*. Group 1 (low)	Reference	0.50 (0.13, 2.03)	0.89 (0.22, 3.67)	**3.37 (1.20, 9.46)**
Trend Across Four Dose Groups	Linear trend p = 0.016 **OR = 1.59 (1.01, 2.31) for each increase in category**			

Notes:

LOD = Limit of Detection; OR = Odds Ratio.

(a) Crude ORs for "Psychomotor Delay" (*i.e.* PDI score ≤ 85) were calculated using counts of subjects in Table 1. Statistically signification associations at a 95% confidence level are highlighted in bold.

(b) The categorization scheme used by Rauh *et al*. (2006).

Table 4 Sensitivity Analysis of Outcome Misclassification for CBCL-Derived Outcomes in Rauh *et al.* (2006)[a] Mothers of children with high prenatal exposure to chlorpyrifos may be more likely to report symptoms on the subjective CBCL survey. We recalculated ORs for CBCL-derived outcomes based on a scenario in which 1, 2 or 3 children in the high exposure category were misclassified as having each outcome. Our results show that misclassification of a small number of exposed subjects strongly attenuates the magnitude of associations.

Outcome	**Reported ORs**		**Crude ORs Calculated Assuming 1, 2, or 3 Exposed Subjects Misclassified with a "Problem"**		
	Crude	**Adjusted**	**1 Subject**	**2 Subjects**	**3 Subjects**
Attention Problems	**11.31 (1.75, 120.89)**	**11.26 (1.79, 70.99)**	**8.83 (1.20, 99.31)**	6.46 (0.71, 78.73)	4.20 (0.29, 59.07)
ADHD Problems	**5.59 (1.14, 29.20)**	**6.50 (1.09, 38.69)**	4.36 (0.77, 24.26)	3.20 (0.45, 19.54)	2.08 (0.18, 15.01)
PDD Problems	2.45 (0.50, 10.15)	**5.39 (1.21, 24.11)**	1.80 (0.29, 8.26)	2.05 (0.18, 14.76)	1.00 (0.20, 10.44)

Notes:

ADHD = Attention Deficit Hyperactivity Disorder; CBCL = Child Behavior Checklist; OR = Odds Ratio; PDD = Pervasive Developmental Disorder.

(a) The reported crude and adjusted ORs for behavioral outcomes associated with high vs. low chlorpyrifos exposure were presented in Rauh *et al.* (2006). The recalculated crude ORs were determined based on an assumption that one or more children in the high exposure category were misclassified as having a "CBCL-related problem." Statistically signification associations at a 95% confidence level are highlighted in bold.